

Comprehensive Reliability

Successful data centre electrical system design involves more than just the number of “9s”

By Robert Weber, Senior Vice President, Environmental Systems Design, Chicago -- Consulting-Specifying Engineer, 11/1/2006

The ability of any mission-critical facility to function properly depends on the sum of its parts. Facility owners need to optimize what they're trying to achieve by balancing their needs against their budget. Each aspect of a data centre's design needs to be appraised, from its smallest component to its major M/E/P systems. Their degree of integration will determine the data centre's “abilities”: availability and reliability, scalability and flexibility, and maintainability.

Availability and reliability

Availability, or the amount of time a system is capable of performing its intended functions, is one of the major measuring sticks in determining the classification of a data centre (see “Data Center Benchmarking,” p. 52). Expressed as a percentage and measured in “9s”—e.g., 99.9999% reliability equals six 9s—availability is determined by reliability, or the equipment failure rate, and the amount of time it takes a system to recover from a failure. As the driving force behind most critical facility's design, reliability is often achieved by designing redundancy into the critical systems at all levels. However, a data centre with more 9s of reliability doesn't necessarily make it more available. Factors such as human error, single points of failure, maintenance, testing and inadequate procedures and processes can severely affect the uptime of a data centre's electrical infrastructure.

To create a data centre with optimal availability and reliability, designers need to understand the implications of human error, which is the cause of more than 15% of all unplanned downtime, according to IT research and analysis firm Gartner Research, Stamford, Conn. Statistical data indicates that the expected failure rate for a team of operators performing a well-designed task is 10-6 failures per hour, the same as the failure rate of some critical equipment itself. However, this increases dramatically when operators are introduced to higher stress levels, with probabilities reaching 0.3 failures per hour. Data centre designers can account for this by making sure human interfaces are clear and well-defined and systems operation isn't over-complicated.

A case in point is the Emergency Power Off (EPO) system. Required by the National Electrical Code, an EPO immediately disconnects all electronic equipment upon activation. While an important safety device, this system also poses a big risk, and in fact, the EPO system is a data centre's single greatest site of human error. Operators can inadvertently touch or hit these switches and shut down the whole system, causing an unintentional loss of electrical power to the critical load.

Standard upgrades for EPO systems, which can enhance availability and reliability, include device covers, local alarms, clear labelling, consistency of location, testing capabilities and double-action devices, which require the operator to perform more than one function at a time to activate the switch. Other common human errors within a data centre include improper overcurrent coordination, poor training and insufficient live testing of components and systems.

Operator training and system testing is also vital to a critical facility's availability. When workers understand the limitations of their systems and know how to properly manage them in emergency situations, the margin of human error can decrease significantly. Facilities can even take advantage of system-simulation software now being developed by equipment manufacturers. These programs model installed systems, enabling operators to experience various conditions, including failed components, and then practice the appropriate response. Although training and testing potentially optimize uptime, many mission-critical facilities lack the confidence in their systems to implement these safeguards and fear interrupting their critical load even to ensure its future continuity.

Another way to significantly increase system availability and reliability is to eliminate or move all single points of failure as close to the computer load as possible. Single points of failure are components or systems that, when failing or turned off, will result in a total loss of critical functions. According to a survey by the Uptime Institute, Santa Fe, N.M., the majority of site infrastructure failures occur between the UPS output and the computer load.

Understanding the interdependencies among the systems within a data centre can also afford designers the ability to lay out the facility more effectively. Knowing that mechanical systems rely on the electrical systems for power and the electrical systems depend on mechanical systems for cooling can change the way a critical facility is organized. Physically segregating these systems from each other is another way to enhance a data centre's reliability. In the event of a fire or other localized disaster, separate electrical components or systems will ideally allow one of the two to continue functioning.

Another consideration with reliability is energy efficiency. In the past, the two were usually thought of as conflicting objectives. However, with rising cost of energy, which has doubled in some sectors due to deregulation, and the escalating power demands of today's data centre's, owners can no longer afford to overlook more sustainable options.

Current studies suggest that more than 50% of the power consumption in a critical facility goes to support IT equipment; even the slightest increase in server efficiency will have a major effect on overall data centre efficiency.

Although the task seems overwhelming, designers can reduce the energy consumed by these systems while still providing the availability and reliability their clients demand. Here's how: In an electrical distribution system, redundancy for the sake of reliability can decrease overall system efficiency. For example, a facility may employ a system-plus-system configuration that loads each of its UPS systems at only 30% to 35%, resulting in low UPS efficiencies. However, by designing a more scalable infrastructure, where the electrical distribution system is intended for the current load, but has the capacity to expand as the demand increases, a data centre can save on energy and still maintain the desired availability.

Most electrical inefficiencies are the result of the facility's cooling system. For example, density variation, where the data centre's overall cooling set point is lowered to account for hot spots in certain areas, has been an operator's rule of thumb for years. Although this does help to cool these spots, it also supplies colder air than necessary to the remainder of the facility. Here, designers can consider a variety of creative strategies that cool each area.

Over-sizing electrical equipment, in which unrealistic load requirements are being employed, can be another major waste of power. Data centre's are designed to bear specific load parameters, which are often based on nameplate-published loads that are typically higher than actual values. Instead, thermal reports produced by individual manufacturers may uncover more precise data on the equipment's realistic power figures.

The sustainability of data centre's is becoming a greater concern to regulators and equipment manufacturers alike. The U.S. Environmental Protection Agency held a conference entitled "Enterprise Servers and Data Centre's: Opportunities for Energy Savings" in January to address the issue. Additionally, major equipment suppliers and a group of technology leaders recently joined forces to develop Green Grid, a not-for-profit association that promotes the reduction of power and cooling demands within data centre's.

Finally, in order to validate whether a system design theoretically meets a data centre's needs, designers and owners alike are demanding the use of reliability modelling, which uses statistical analysis to predict the failure rate of systems and components, more accurately determining a facility's potential availability.

Scalability and flexibility

With today's ever-changing business landscape and the evolution of new technology, there is a growing demand for critical facilities to be both scalable and flexible. A scalable system gives a data centre owner the opportunity to purchase only the equipment required to support present-day needs, with the ability to add components at a later date without affecting critical operations. This helps reduce initial capital costs and allows systems to operate closer to their full-load ratings, thus increasing efficiency. Here's a classic example: Our firm recently designed a corporate data centre intending initially only to use a portion of its raised floor area. The facility's main services and switchgear were planned to sustain its full future load, with only sufficient UPS, standby generator and raised-floor distribution capacity installed to support the initial load, plus some predetermined growth. The addition of future components were accounted for in the initial design, and would not require an interruption to the facility or be a risk to its critical loads.

Another way scalability is being addressed is with modular design. Here, the critical facility's raised-floor area is designed in modules or zones to accommodate a designated amount of equipment, with the ability to increase down the road.

A module typically consists of an independent, raised-floor area and can be defined by the number of racks or rows of equipment grouped together. Modules can be constructed to serve various functions and are built out in increments to provide flexibility, allowing data centre's to take advantage of economy of scale and new technologies.

Currently, ESD is working on a data centre for a client where the modules are based on 1,000 sq. ft. Some of these modules are designed for high-density loads in excess of 100 watts per sq. ft., while others are less than 50 watts per sq. ft. in order to give the client more flexibility should its loads change, densities increase or technology needs change. The central electrical and mechanical plant loads are also being sized to support an average load over all the modules.

Maintainability

In years past, maintenance personnel had windows of opportunity to perform scheduled maintenance on a data centre, typically during holidays or designated off weekends. But, today's high-availability facilities do not have such a luxury. Data centre's must now be concurrently maintainable. This means regular maintenance and infrastructure upgrades need to be performed without shutting down the facility's critical load, a task that needs to be considered when designing both the electrical and mechanical systems.

It is for this reason that redundancy is crucial to data centre maintainability. A redundant electrical infrastructure gives operators and facilities personnel the opportunity to complete required maintenance, thereby creating a more reliable data centre.

One of the best ways to encourage redundancy is to design a data centre with two active power paths (2N), or a system-plus-system design. This configuration typically consists of two independent UPS systems (Sides A and B in Figure 1, p. 52), each supplying power to its own path, but also capable of supporting the total critical distribution load. Using this concept, if one side fails or requires maintenance, its load will subsequently transfer over to the other path. In this case, one side can be serviced at a time while still maintaining continuous uptime.

Touching briefly again on reliability, the 2N system also helps in that area. When dual-corded equipment is used in conjunction with the dual-path system described above, there are no single points of failure down to the plug level. The single point of failure for data centre's with single-corded equipment is the static transfer switch. By locating this switch on the secondary side of the PDU transformer or preferably at the rack level, availability can be significantly increased as well.

Designing a data centre to meet the availability requirements of today's mission critical facilities, while still being nimble enough to adapt to changing technology and an owner's business requirements, is very challenging. While no one design works for all scenarios, it is critical that a data centre designer successfully integrates the core design ingredients mentioned above into the overall design.

Data Center Benchmarking

A modern, state-of-the-art data centre can be characterized by its level of availability. The industry benchmark used to define this is called a "tier." Developed 10 years ago by the Uptime Institute, Santa Fe, N.M., and updated in 2006, tier classifications are as follows:

Tier 1. Single path power and cooling distribution, without redundant components, providing 99.67% availability.

Tier 2. Single path for power and cooling distribution, with redundant components, providing 99.75%.

Tier 3. Multiple active power and cooling distribution paths—with only one active—redundant components, and is concurrently maintainable, providing 99.98%.

Tier 4. Composed of multiple active power and cooling distribution paths, has redundant components and is fault tolerant, providing 99.99%.

Additionally, the Telecommunication Industry Assn. (TIA) published Standard TIA-942 "Telecommunication Infrastructure Standard for Data Centre's" in 2005, which expands the definition of the above tier levels. Subsequently, many design firms and individual corporations have developed their own classifications or tier structures, helping them to increase consistency company-wide.

Modern state-of-the-art data centre's are made up of many independent components and systems that are designed to a variety of tier levels, as reflective of the budget or individual needs of the businesses they support. This may result in a different overall tier classification for the larger data centre. Visit www.upsite.com for more information.

Arc Flash: a Growing Hazard

Along with the increased power densities of today's data centre's another electrical issue designers must account for is arc flash—current that passes through the air when the insulation or isolation between electrified conductors is no longer sufficient to withstand the applied voltage.

An arc flash can lead to severe burns for unprotected personnel. According to the NFPA, more than 2,000 people are treated annually from such instances. Therefore, a classification system has been created to quantify the danger in working with each piece of equipment. Designers are asked to label machinery warning operators to wear appropriate clothing and provide tips on how to maintain and service the equipment. But the arc flash phenomenon has created another dilemma in the design of critical distribution systems as arc-flash safety and overcurrent coordination in many cases seem to be competing issues.

From an arc-flash standpoint, the faster an arcing fault is detected and cleared from the system, the less energy it releases into the air. For many years, designers have either not installed instantaneous trip functions, or set them high in order to increase breaker coordination, thus delaying the time it takes the breaker to trip. Given the current focus on arc-flash safety, it is critical that designers understand the characteristics of the overcurrent protective devices they specify and select those that will provide the best combination of protection for both the worker and the electrical system.